데이터의 한계를 넘어, 비즈니스의 미래를 연산하다

AI 인프라부터 플랫폼, 서비스까지
AI Infra 오케스트레이터

AI Infra Platform by ITCEN CTS

"차세대 AI/HPC 솔루션으로 복잡한 시뮬레이션부터 AI 모델링까지 압도적인 성능을 경험하세요"

1 복잡한 AI 자원을 한눈에 : AI Infra Management

이기종 AI/HPC 자원의 통합 가시성

GPU, CPU, 고속 네트워크, 스토리지 등 분산된 HPC 및 AI 인프라 자원을 하나의 화면에서 직관적으로 파악
복잡한 자원 상태를 단순한 인사이트로 전환

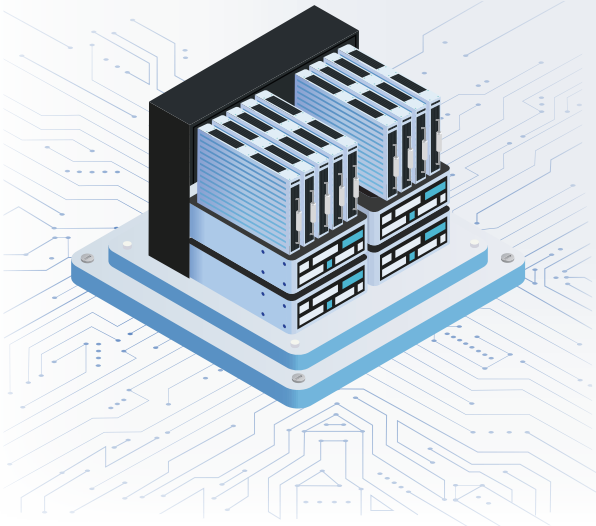
워크로드 중심의 유연한 자원 활용

AI 학습, 시뮬레이션, 대규모 연산 워크로드 특성에 따라 자원 활용 현황을 분석하고 효율적인 사용을 지원
변화하는 연산 요구에 민첩하게 대응할 수 있는 인프라를 지향

대규모 연산 환경을 위한 지능형 운영 관리

안정적인 AI · 과학 연산 운영 기반 제공

2 고속 연산과 데이터 처리에 최적화된 AI/HPC 인프라



초고속 대역폭 컴퓨팅 노드

NVIDIA. HPE

- **GPU Server** : 안정성과 성능이 검증된 최신 GPU 서버 구성
- **Slicing Tech** : GPU의 물리적 자원을 효율적으로 분배하는 Fine-tuning 및 GPU Passthrough 기술적용

초저지연 스토리지

VAST DELL Everpure

- **Data Platform** : 오브젝트·파일 기반을 아우르는 통합 스토리지 플랫폼
- **AI Optimized I/O** : 대규모 모델 학습 워크로드에 최적화된 고대역폭/저지연 처리

무손실 네트워크

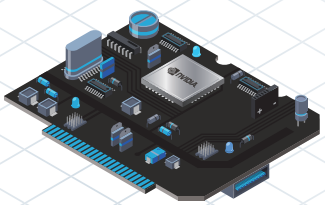
NVIDIA. HPE ARISTA CISCO

- **Lossless Ethernet :** RoCE 기반 고성능 네트워크 구현
- **Low Latency :** 초저지연 통신으로 GPU 간 데이터 전달 최적화
- **High Throughput :** AI/HPC 워크로드에 최적화된 고대역폭 제공

One-Stop Appliance

- Compute–Network–Storage를 통합한 단일 어플라이언스로 AI/HPC 인프라 구축의 복잡성을 제거하고 운영 효율을 극대화
- AI Server, Nvidia GPU, Storage가 통합된 단일 패키지 형태로 제공하여 구축 복잡성 제거

3 초거대 AI를 위한 혈관: 고성능 AI 네트워크 아키텍처



고성능 High Speed Network(InfiniBand & RoCEv2)

- **Infiniband** : 전용 RDMA 패브릭을 기반으로 AI·HPC 워크로드에서 최고 수준의 초저지연·고대역폭 성능을 제공
- **RoCEv2** : 이더넷 기반 RDMA로 기존 네트워크 인프라와 자연스럽게 통합되며 AI 환경에서도 안정적인 저지연 성능을 제공

The Specs

High Speed	400G/800G/1.6 Tbps 초고속 초지연의 무손실 네트워크
Compatibility	다양한 벤더의 GPU·네트워크 장비와 완전한 인터페이스 호환성을 확보해 유연한 AI 클러스터 구성을 지원
Latency	AI 워크로드에 최적화된 초저지연 트래픽 처리

AI Infra Management

솔루션 제공 기능

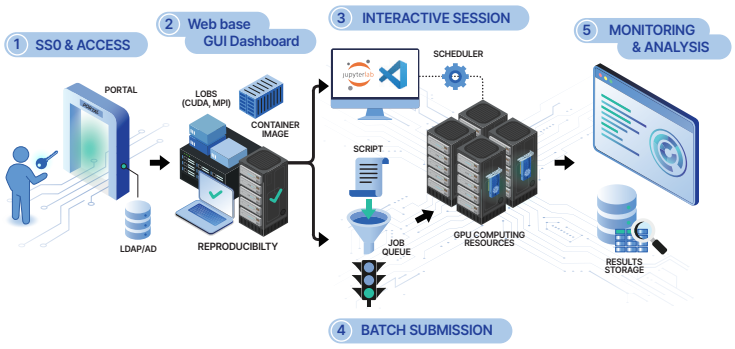
관리효율
프로젝트 중심의 강력한 거버넌스 통제

사용편의
직관적인 웹 기반 셀프 서비스 환경

자원 최적화
지능형 스케줄링 기반 자원 효율 극대화

운영 통찰
데이터 기반의 실시간 인프라 인사이트

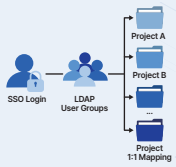
표준 HPC 워크플로우 지원



1. 통합 인증 및 접속 (Single Sign-On & Access)
2. 웹 브라우저 기반의 GUI 대시보드
3. 대화형 분석 및 코딩(Interactive Session)
4. 작업 제출 및 스케줄링(Monitoring & Analysis)
5. Job & GPU 모니터링

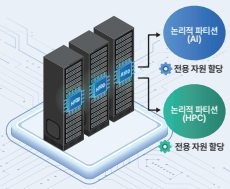
거버넌스 기반의 자원 최적화

통합 신원 및 권한 관리 (SSO&IAM)

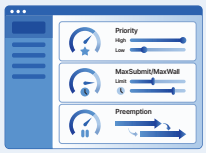


중앙 집중식 통제

전략적 자원 파티셔닝 (Partition Management)



유연한 자원 소비 정책 (Advanced Qos Settings)



SLA 준수

- **통합 인증 및 접속 (Single Sign-On & Access)**
  #보안접근 #계정자동동기화 #Role_Based_Access_Control
- **전략적 자원 파티셔닝(Partition Management)**
  #전용노드배정 #물리자원격리 #효율적_인프라_운용
- **유연한 자원 소비 정책 (Advanced QoS Settings)**
  #자원독점방지 #우선순위스케줄링 #SLA준수

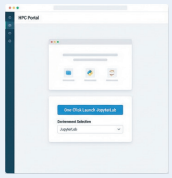
투명한 자원 활용 및 데이터 인사이트

프로젝트 기반 자원 할당량 관리 (Quota & Account)

프로젝트별로 할당된 컴퓨팅 시간(Tflops/h) 및 GPU 사용량 실시간 확인. 예산 기반의 자원 관리로 계획적인 연구 수행 지원.

#쿼터관리 #프로젝트별_비용추적 #투명한_자원배분

사용자 친화적 셀프 서비스



복잡한 정책 설정은 뒤로하고, 연구원은 CTS 인프라 관리포털을 통해 할당받은 자원 내에서 클릭 한 번으로 모듈 환경 로드 및 JupyterLab 실행.

#클릭기반_HPC #개발환경_즉시구동 #사용자_편의성

실시간 모니터링 및 활용도 분석 (Visibility)

전체 클러스터의 가동률, 프로젝트별 GPU 점유율, 사용자별 Job 성공률 등을 시각화된 리포트로 제공. 차년도 인프라 도입을 위한 데이터 기반의 인사이트 제공.

#GPU사용량통계 #대시보드 #데이터기반_의사결정

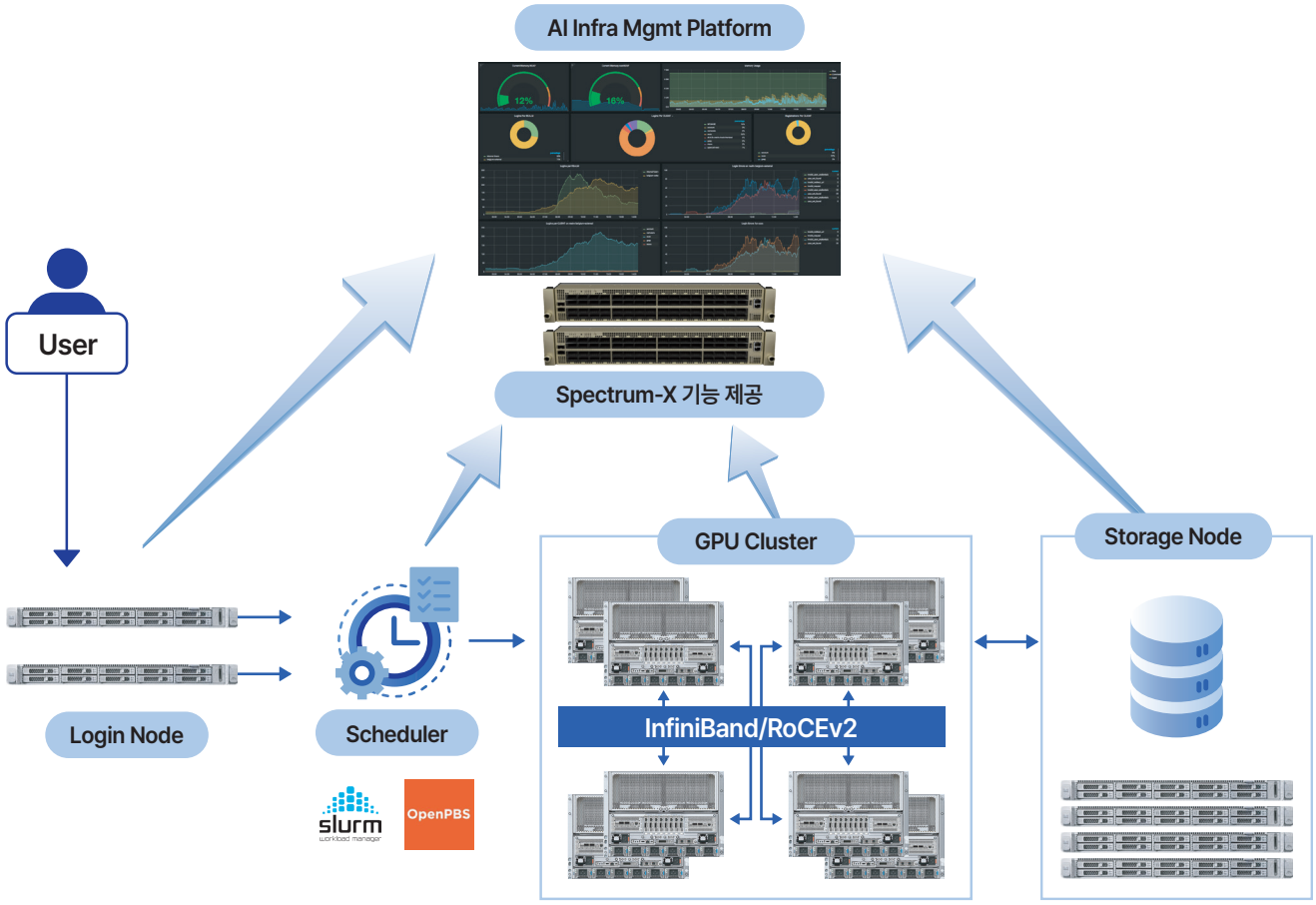
- **프로젝트 기반 자원 할당량 관리 (Quota & Account)**
  #쿼터관리 #프로젝트별_비용추적 #투명한_자원배분
- **사용자 친화적 셀프 서비스**
  #클릭기반_HPC #개발환경_즉시구동 #사용자_편의성
- **실시간 모니터링 및 활용도 분석 (Visibility)**
  #GPU사용량통계 #대시보드 #데이터기반_의사결정

AI/HPC 인프라

최신 GPU 및 고속 인터커넥트 기술 적용
Key Features

컴퓨팅 노드: 최신 아키텍처 기반의 고집적 GPU 서버 구성
스토리지: 병렬 파일 시스템(Lustre, GPFS 등) 기반의 초고속 I/O 처리
네트워크: 초저지연 RDMA를 지원하는 InfiniBand·RoCEv2 네트워크

AI Infra Full-Stack Architecture



AI/HPC 필수 인프라 스택

초고속 대역폭
컴퓨팅 노드

- 최신 아키텍처 기반 GPU 서버 노드로 대규모 모델 학습·추론 수행
- 멀티 GPU·멀티 노드 분산 학습 구성을 위한 고속 인터커넥트 지원
- GPU Slicing·Passthrough·가상화 환경 지원으로 다양한 워크로드 요구 대응

초저지연
스토리지

- GDS(GPU Direct Storage) 지원 스토리지 구성
- I/O 병목 최소화 병렬파일 시스템(NFS-RDMA, Lustre, GPFS 등) 기반의 고성능 데이터 처리
- Object + File 기반의 통합 Data Platform으로 수집→전처리→학습 전 과정 지원

무손실 네트워크
네트워크

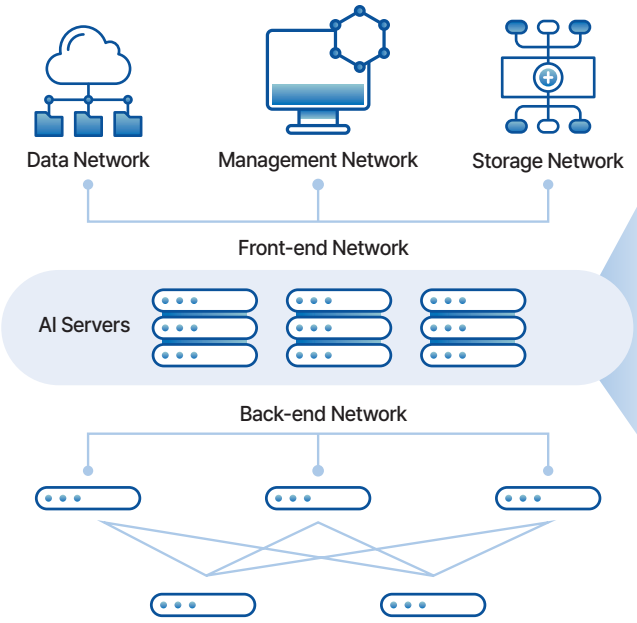
- InfiniBand·RoCEv2 기반 RDMA 패브릭으로 GPU 간 고속 데이터 전송을 지원
- Leaf–Spine 구조의 고대역폭 네트워크로 대규모 클러스터 확장 가능
- PFC·ECN 기반 무손실 패브릭 구성으로 학습/추론 트래픽의 안정성 확보

고성능 AI 네트워크 아키텍처

AI/HPC Infra 네트워크 요구사항

- 초저지연
- 무손실
- 대규모 확장성
- 예측 가능한 성능

AI 네트워크 구성을 위한 핵심 요소



1 Non-Blocking 네트워크 구성
- 대량의 AI 학습 트래픽은 작은 병목도 전체 학습 시간에 지연 발생
- 네트워크 전 구간에서 병목이 발생하지 않도록 대역폭을 100% 활용 가능한 구조 필수

2 800G/400G/1.8Tbps 기반
- 최신 GPU는 노드당 수백 Gbps의 통신량을 요구
- 백엔드 패브릭이 400G~800G, 1.6Tbps 수준으로 업그레이드되어야 GPU 성능을 온전히 활용 가능

3 Low latency 네트워크
- AI Distributed Training에서 미세 지연 차이가 전체 학습 시간에 직접 영향을 미침
- RDMA를 통한 초저지연 네트워크 필요

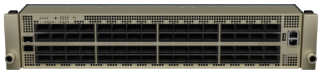
4 Lossless 네트워크
- AI 통신은 특성상 패킷 하나만 손실되어도 전체 GPU가 다시 동기화를 기다려야 하는 구조
- 패킷 Drop이 0에 가깝도록 보장 필요

5 혼잡제어(Congestion Control)
- AI 트래픽은 순간적인 Microburst가 자주 발생
- 정교한 혼잡 탐지·완화·속도 제어 필요

RoCE 주요 제품

NVIDIA

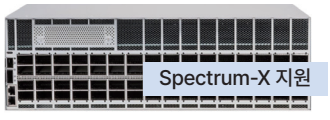
NVIDIA Spectrum SN4700
32 × 400Gb/s



NVIDIA Spectrum SN5610
64 × 800Gb/s

NVIDIA Spectrum SN6810
128 × 800Gb/s

CISCO



Cisco N9364E-SP2R
64 × 800Gb/s

Cisco N9K-C9364D-GX2A
64 × 400Gb/s

Cisco N9K-C9348D-GX2A
48 × 400Gb/s

데이터의 한계를 넘어,
비즈니스의 미래를 연산하다

핵심 비즈니스 영역

ITCEN Group 파트너십을 통한
AI 인프라 유연성과 비용 최적화 동시 확보

H/W Partner

HPE · IBM · ARISTA · CISCO · SUPERMICRO · DELL EMC · VAST · NVIDIA

Cloud
유연한 확장과 신속한 구축

Hybrid
탄력적 구조와 선택적 구성

On-premise
안정적 운영과 강화된 보안

ITCEN Public Cloud

Advanced	Premier
aws	Google Cloud
Gold Partner	Premier
Microsoft Azure	Naver Cloud
Platinum	Silver
NHN CLOUD	kt cloud

ITCEN CTS

(주)아이티센씨티에스 www.itcencts.com
TEL.02-3289-0114 FAX. 02-836-6611 Email. ai.cts@itcen.com
경기 과천시 과천대로12길 117, 과천 펜타원 E동(아이티센타워)

제품 사양은 2025년 10월 기준이며, 고지 없이 변경될 수 있습니다.
관련 문의는 당사 홈페이지의 또는 영업대표를 통해 확인 바랍니다.